

Е.И. Козлова, Д.М. Клименко
**ОБНАРУЖЕНИЕ СОСТЯЗАТЕЛЬНЫХ АТАК ЧЕРЕЗ
ВЫЯВЛЕНИЕ АНОМАЛИЙ В СКРЫТЫХ СЛОЯХ НЕЙРОСЕТИ**

Козлова Е.И., Клименко Д.М. Обнаружение состязательных атак через выявление аномалий в скрытых слоях нейросети.

Аннотация. Современный уровень развития информационных технологий характеризуется широким использованием методов искусственного интеллекта (ИИ), особенно глубокого машинного обучения, в таких важных областях, как здравоохранение, транспорт, финансовые услуги, системы биометрической аутентификации и национальная безопасность. Модели нейронных сетей все чаще используются для принятия решений, которые напрямую влияют на безопасность, здоровье и финансовые интересы отдельных лиц и организаций. Однако по мере повышения функциональной значимости систем искусственного интеллекта растет и их уязвимость к целенаправленным атакам. Существующие подходы к противодействию, такие как состязательное обучение или входная фильтрация, обладают существенными ограничениями: они либо снижают точность модели на легитимных данных, либо эффективны лишь против заранее известных типов атак, не обеспечивая обнаружения новых угроз. В этих условиях возрастает потребность в механизмах детекции, способных определить сам факт атаки независимо от ее реализации. При этом различные исследования показывают, что состязательные примеры вызывают аномальные паттерны активаций в скрытых слоях нейронной сети, качественно отличающиеся от тех, что наблюдаются при обработке легитимных данных. Это открывает возможность применения методов обнаружения аномалий, таких как изоляционный лес или расстояние Махаланобиса – для выявления подозрительных запросов без модификации исходной модели. Особую значимость приобретает также влияние архитектурных и обучающих гиперпараметров на внутренние представления модели и, как следствие, на ее уязвимость и обнаружение атак. Целью данной работы является экспериментальное исследование метода обнаружения состязательных атак на основе анализа аномалий в активациях скрытых слоев нейронной сети.

Ключевые слова: искусственный интеллект, глубокое машинное обучение, нейронные сети, состязательная атака, обнаружение атак, аномалии, скрытые активации, внутренние представления модели, изоляционный лес, метрики оценки качества модели.

1. Введение. Состязательные атаки (англ. Adversarial attacks) представляют собой целенаправленное воздействие на входные данные нейросетевой модели с целью вызвать ее ошибочное поведение при сохранении визуальной неизменности исходного объекта для человека. Эти методы злонамеренного изменения входных данных представляют особую угрозу, поскольку при их воздействии визуально незаметные (для человека) возмущения приводят к грубым ошибкам классификации [1, 2]. Такие атаки демонстрируют фундаментальное различие между восприятием человеком данных и их интерпретацией в модели нейронной сети и могут быть использованы для обхода систем контроля доступа, искажения

медицинских диагнозов или дезинформации [2]. В условиях, когда ИИ интегрирован с системами контроля доступа, медицинской диагностики или автономного управления, успешная атака противника может иметь серьезные последствия, включая потерю личных данных, финансовые потери или опасные для жизни ситуации [3, 4]. Исследования показали, что добавление едва различимого шума к изображению может привести к полной потере точности сверточной нейросети. Это открытие выявило фундаментальную уязвимость глубоких моделей, обусловленную их высокой нелинейностью и чувствительностью к малым возмущениям в областях с низкой плотностью обучающих данных [5]. Универсальность феномена состязательных возмущений проявляется в его применимости ко всем основным типам данных, обрабатываемых современными моделями машинного обучения. В данном исследовании атаки проводятся на изображениях [6].

Современная классификация состязательных атак строится по двум основным критериям:

1. Уровень доступа атакующего к модели:
 - Белый ящик (White-box) – атакующий имеет полный доступ к архитектуре, параметрам и градиентам модели.
 - Черный ящик (Black-box) – атакующий взаимодействует с моделью только через вход/выход, не имея информации о внутреннем устройстве [7 – 9].
2. Целевое поведение:
 - Целевые (targeted) – атака направлена на принуждение модели к предсказанию конкретного неверного класса.
 - Нецелевые (untargeted) – достаточно вызвать любую ошибку классификации [10].

Эти категории определяют сложность и реализуемость атак в практических сценариях, особенно в системах информационной безопасности. Так, в случае «белого ящика» атаки, несмотря на их высокую эффективность, редко реализуемы в условиях развернутых коммерческих или государственных ИИ-систем, где внутренняя архитектура модели скрыта за интерфейсом прикладного программирования (API) и защищена механизмами обфускации или инкапсуляции. Напротив, сценарии «черного ящика» отражают реалистичные условия противодействия, в которых злоумышленник ограничен возможностью отправки запросов и анализа ответов, что делает такие атаки более релевантными для оценки операционной устойчивости систем [8]. Аналогично, целевые атаки требуют большего вычислительного бюджета и точного знания пространства выходных

меток, однако представляют наибольшую угрозу в сценариях, связанных с подделкой биометрической идентификации или манипуляцией диагностическими системами, где требуется достижение конкретного ложного результата [10]. В то же время нецелевые атаки, будучи менее ресурсоемкими, достаточны для нарушения целостности таких систем, как фильтрация контента, антивирусная проверка или автоматизированная оценка модели, где любая ошибка классификации может привести к нарушению бизнес-логики или утечке данных. Таким образом, комбинированное рассмотрение уровня доступа и типа цели позволяет строить адекватные модели угроз и разрабатывать методы защиты, соответствующие конкретному классу рисков.

На рисунке 1 представлен пример данных, содержащих как «чистые» изображения (x – изображение исходное, z – изображения исходные с проведенными на них состязательными атаками), так и преобразованные для проведения состязательных атак ($G(x)$ – исходное изображение с примененным к нему состязательным обучением, $G(z)$ – предобученные изображения после состязательных атак) [11].

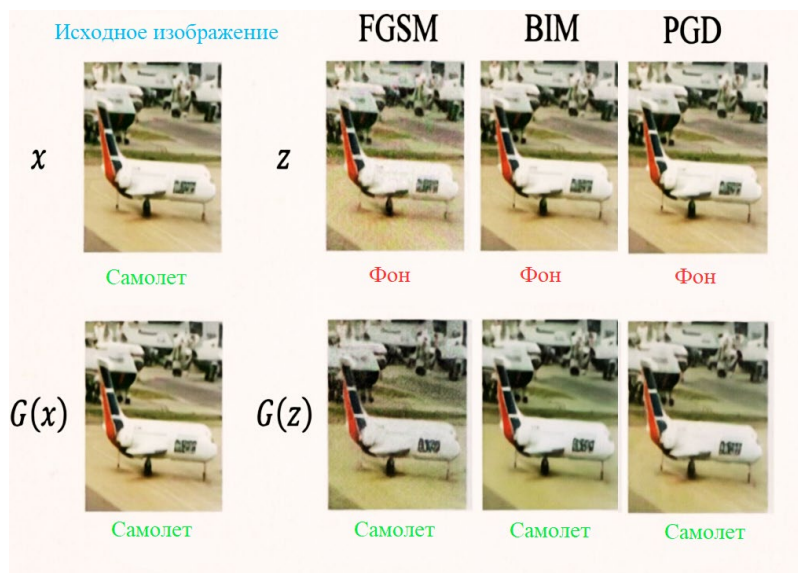


Рис. 1. Влияние различных типов атак (FGSM, BIM, PGD) на классификацию объекта «Самолет»

Целью данной работы является экспериментальное исследование эффективности метода обнаружения различных состязательных атак на основе анализа аномалий в активациях скрытых слоев нейронной сети. Для достижения цели поставлены следующие задачи:

1. Провести обзор существующих методов обнаружения состязательных воздействий.
2. Реализовать последовательность этапов генерации атак и извлечения промежуточных активаций.
3. Обучить и протестировать детектор аномалий на основе методов машинного обучения.
4. Оценить эффективность подхода с использованием метрик площади под кривой (AUC) и доли истинно положительных результатов (TPR) при фиксированном значении доли ложноположительных результатов (FPR).
5. Исследовать влияние гиперпараметров модели на способность к обнаружению атак.

2. Обзор методов генерации состязательных атак. Формально, состязательный пример x_{adv} для исходного объекта $x \in R^d$ с истинной меткой y определяется как $x_{adv} = x + \delta$, где возмущение $\delta \in R^d$ удовлетворяет условию $\|\delta\|_p \leq \epsilon$ при малом $\epsilon > 0$ и приводит к нарушению корректности предсказания: $f_\theta(x_{adv}) \neq y$. При этом визуальное восприятие x и x_{adv} человеком остается идентичным, что подтверждается психофизиологическими исследованиями порогов различимости изображений [6].

Современные методы генерации состязательных примеров преимущественно опираются на градиентную информацию функции потерь модели, что позволяет целенаправленно формировать минимальные по норме возмущения, способные нарушить корректность предсказания. Среди множества существующих методов наибольшее распространение получили Метод быстрого градиентного знака (Fast Gradient Sign Method, FGSM), Проецируемый спуск по градиенту (Projected Gradient Descent, PGD) и Базовый итеративный метод (Basic Iterative Method, BIM), отличающиеся по вычислительной сложности, точности аппроксимации и устойчивости к защитным механизмам [12].

Метод быстрого градиентного знака (FGSM), введенный в работе Goodfellow et al. [11], основывается на линейной аппроксимации функции потерь $L(f_\theta(x), y)$ в окрестности входного вектора (исходное изображение) $x \in [0, 1]^d$. Состязательный пример формируется по формуле:

$$x_{adv} = \prod_{[0,1]^d} (x + \varepsilon \cdot \text{sign}(\nabla_x L(f_\theta(x), y))),$$

где $\varepsilon > 0$ – параметр, ограничивающий L_∞ – норму возмущения, а $\prod$ – оператор проекции на допустимый диапазон пикселей. Несмотря на вычислительную эффективность (один прямой и один обратный проход), FGSM часто генерирует возмущения, не являющиеся оптимальными, особенно против моделей с гладкими градиентами, поскольку игнорирует нелинейную структуру функции потерь [7].

Проецируемый спуск по градиенту (PGD) устраняет этот недостаток за счет итеративного уточнения направления возмущения [7]. Формально метод PGD реализует многократный градиентный подъем с проекцией на L_∞ – шар радиуса ε :

$$x_{t+1} = \prod_{B_\varepsilon^\infty(x)} (x_t + \alpha \cdot \text{sign}(\nabla_x L(f_\theta(x_t), y))),$$

где α – шаг итерации, $B_\varepsilon^\infty(x) = \{x' \in R^d : \|x' - x\|_\infty \leq \varepsilon\}$. При малом α и достаточном числе шагов метода PGD аппроксимирует локальный максимум функции потерь и считается «сильнейшей» L_∞ – ограниченной атакой в сценарии белого ящика [7]. Например, на модели ResNet-18, обученной на CIFAR-10, PGD с $\varepsilon = \frac{8}{255}$, $\alpha = \frac{8}{255}$, $T = 20$ снижает точность с 92% до менее 5% [13].

Базовый итеративный метод (BIM) формально совпадает с методом PGD, однако в оригинальной постановке применяется без случайного старта ($x_0 = x$) и часто интерпретируется как итеративная версия метода FGSM. Метод BIM обеспечивает баланс между вычислительной эффективностью и качеством возмущения, демонстрируя более высокий уровень успеха по сравнению с методом FGSM при сопоставимых затратах [14, 15].

3. Построение детектора аномалий. Построение детектора аномалий в настоящем исследовании выполнено в соответствии с парадигмой обучения без учителя [16]. Для выявления состязательных примеров использовалась стратегия построения детекторов аномалий с обучением без учителя, основанных на анализе внутренних представлений, формируемых в промежуточных слоях предварительно обученной сверточной нейронной сети [17]. Обучение детектора проводилось только на выборке чистых (неискаженных) данных, что соответствует сценарию обучения на одном классе [18], при котором нормальное поведение модели считается единственным доступным классом.

В настоящее время существует большое число методов обнаружения аномалий [2]. В эксперименте были реализованы и исследованы четыре из них:

1. Изоляционный лес (Isolation Forest, IF) – ансамблевый метод, основанный на случайном разбиении пространства признаков, где аномалии идентифицируются по меньшей средней длине пути до листа в дереве изоляции. Гиперпараметры: $n_estimators=100$, $contamination=0.1$.

2. Одноклассовый метод опорных векторов (One-Class Support Vector Machine, OCSVM) – метод, максимизирующий зазор между началом координат и гиперплоскостью, ограждающей область сосредоточения нормальных данных в признаковом пространстве с ядром RBF. Гиперпараметры: $\nu=0.1$, $\gamma='scale'$.

3. Анализ главных компонент (Principal Component Analysis, PCA) – линейный метод снижения размерности; аномальность вычисляется как квадрат ошибки реконструкции после проекции на главное подпространство, содержащее 95 % дисперсии. Порог детекции устанавливается как 90-й перцентиль ошибки реконструкции на обучающей (неискаженной) выборке.

4. Расстояние Махаланобиса (Mahalanobis distance) – статистическая мера удаленности до центра класса с учетом ковариационной структуры признакового пространства; теоретически обоснован как наиболее чувствительный к смещениям, вносимым состязательными возмущениями [2].

Для каждого детектора вычисляются оценки аномальности:

- для IF и OCSVM – отрицательная функция принятия решений (чем выше значение, тем выше вероятность аномалии);
- для PCA – евклидова норма разности между исходным вектором активаций и его реконструкцией.

Качество детекции оценивалось с помощью следующих метрик:

- площадь под ROC-кривой (AUC ROC) – интегральный показатель способности метода располагать аномальные образцы выше нормальных в ранжированном списке [19];
- $TPR@FPR = \tau$ (для $\tau \in \{1\%, 5\%, 10\%\}$) – доля истинно положительных срабатываний при заданном уровне ложноположительных срабатываний.

Выбранные для оценки качества детекции метрики в комплексе отражают как общую способность к разделению классов (AUC ROC), так и практическую применимость в условиях строгих требований к доле ложных срабатываний – критически важному аспекту для систем информационной безопасности:

- интегральная метрика бинарного классификатора, инвариантная к порогу AUC ROC (Area Under the Receiver Operating Characteristic Curve) – позволяет сравнивать методы в общем, но не отражает их поведение при реальных операционных ограничениях;

- метрика $TPR@FPR=X\%$, отражающая долю истинно положительных срабатываний при фиксированной доле ложноположительных срабатываний – ключевая для приложений информационной безопасности, где ложные срабатывания (FPR) недопустимы выше заданного уровня (например, 1% – не более 1 ложного тревожного сигнала на 100 легитимных запросов). В исследовании анализировались три уровня: $FPR = 1\%$, 5% , 10% , что охватывает диапазон от режимов с минимально допустимым риском (ядерные системы, финансовый сектор) до режимов с допустимым уровнем ложных срабатываний (внутренние инструменты мониторинга).

Применение такого комплекса метрик позволяет анализировать эффективность детекторов как в общем, так и при строгих ограничениях на долю ложных срабатываний, что особенно актуально для систем информационной безопасности и значимо для практических задач с жесткими требованиями к избирательности [14]. Выбор именно этих методов также обусловлен их различной вычислительной сложностью, требованиями к данным и совместимостью с промышленными практиками разработки и эксплуатации (DevOps и MLOps):

- IF и OCSVM легко конвертируются в формат для хранения/передачи (сериализуются с помощью `joblib/pickle`) и развертываются в виде отдельных программных сервисов;

- PCA может быть реализован на графическом процессоре или даже в виде оператора формата ONNX, что ускоряет этап вывода;

- Расстояние Махаланобиса требует лишь хранения векторов средних и обратной ковариационной матрицы (либо ее диагональной аппроксимации), что обеспечивает минимальные накладные расходы по памяти (< 1 МБ даже для 1024-мерного пространства).

Таким образом, предложенный подход реализует метод обнаружения аномалий в пространстве признаков [20, 21], при котором уязвимость классификатора к состязательным воздействиям компенсируется внешней системой мониторинга, не требующей изменения архитектуры или повторного обучения целевой модели.

4. Исследование эффективности обнаружения состязательных атак. В работе проведены экспериментальные исследования, направленные на анализ устойчивости сверточных нейронных сетей к состязательным атакам и оценку эффективности

четырёх вышеуказанных методов обнаружения аномалий в признаковом пространстве. Эксперименты проводились на четырёх моделях – Model_A, Model_B, Model_A_AT, Model_B_AT, исследованных авторами ранее [22]. Модели построены на архитектурах сверточных нейронных сетей, отличающихся глубиной, размером ядер и стратегией понижения размерности. В качестве признакового пространства использовались векторные активации заданного сверточного слоя: для модели Model_A – выход второго сверточного блока с операцией субдискретизации, для Model_B – выход третьего сверточного блока. Для моделей A и B также были обучены версии (модели Model_A_AT, Model_B_AT) с использованием состязательного обучения (AT), применяющие атаки на основе проекционного градиентного спуска (PGD) на этапе обучения – один из наиболее эффективных в настоящее время подходов к повышению устойчивости. Активации извлекались без вычисления градиентов, нормализация входных данных сохранялась согласно этапу предобработки.

В качестве тестовой среды выбран набор данных Fashion-MNIST – стандартный эталонный набор для задач компьютерного зрения, представляющий собой коллекцию из 70000 изображений одежды 10 классов (28×28 пикселей, в градациях серого). Выбор именно этого набора данных обусловлен следующими факторами:

1. Достаточная сложность по сравнению с набором MNIST при сохранении высокой воспроизводимости результатов.
2. Отсутствие этических и лицензионных ограничений на использование.
3. Широкая распространенность в исследованиях устойчивости к состязательным атакам (например, в работах [8, 10]), что позволяет сопоставлять полученные результаты с данными научной литературы.

Генерация состязательных примеров осуществлялась с использованием библиотеки torchattacks и включала три типа атак, различающихся по механизму и «заметности»: метод быстрого градиентного знака (Fast Gradient Sign Method, FGSM), проекционный градиентный спуск (Projected Gradient Descent, PGD), базовый итеративный метод (Basic Iterative Method, BIM).

Для каждой атаки тестировались три уровня интенсивности возмущения: $\varepsilon = 0.1, 0.15, 0.2$, что соответствует смещению пикселя на 25.5, 38.25 и 51 единицу яркости (при исходном диапазоне [0,255]), и позволяет исследовать зависимость эффективности детекции от силы атаки.

4.1. Влияние типа состязательной атаки на эффективность обнаружения. Одним из ключевых факторов, определяющих уязвимость систем машинного обучения, являются характер и механизм генерации состязательных возмущений. В рамках данного исследования были протестированы три классических типа атак – FGSM, PGD и BIM, – различающиеся по числу итераций, наличию проекции и степени «заметности» вносимых обнаружений. Анализ результатов позволяет сделать вывод о том, что не все атаки равнозначны с точки зрения обнаружения, и выбор тестового набора атак напрямую влияет на оценку надежности системы защиты.

Атаки FGSM (метод быстрого градиентного спуска) и PGD (проекционный градиентный спуск) относятся к классу градиентных методов с контролируемой L -нормой, где величина возмущения ограничена параметром ε . Несмотря на различия в реализации (FGSM – одношаговая атака, PGD – итеративный метод), обе они демонстрируют сходное поведение с точки зрения обнаружения на незащищенных моделях Model_A, Model_B.

Как видно из таблицы 1, для Model_A при $\varepsilon = 0.1$ разница в метриках не превышает статистически значимого уровня ($\Delta TPR < 1\%$ при $N=500$, N – это объем тестовой выборки, использованной в экспериментах: NUM_TEST = 500), что позволяет рассматривать FGSM и PGD как взаимозаменяемые при оценке на незащищенных моделях.

Однако при переходе к моделям с защитным обучением различия становятся значимыми. Как показано в таблице 2, для Model_A_AT при $\varepsilon = 0.2$ PGD обнаруживается хуже FGSM на 5-7 % по $TPR@FPR = 1\%$.

Таблица 1. Сравнение эффективности обнаружения FGSM и PGD на модели Model_A ($\varepsilon = 0.1$)

Детектор	FGSM (AUC)	PGD (AUC)	FGSM (TPR@FPR = 1%)	PGD (TPR@FPR = 1%)
OCVSM	0.9987	0.9987	0.962	0.968
IF	0.9931	0.9931	0.848	0.860
PCA	1.0000	1.0000	0.000	0.000

Таблица 2. Сравнение $TPR@FPR = 1\%$ для FGSM и PGD на модели Model_A_AT ($\varepsilon = 0.2$)

Детектор	FGSM	PGD	Разность (FGSM - PGD)
Mahalanobis	0.914	0.860	+0.054
PCA	0.768	0.696	+0.072

Это объясняется тем, что защитное обучение, выполненное с использованием PGD, формирует устойчивость именно к данному типу возмущения, в то время как FGSM остается относительно незнакомой атакой – что согласуется с выводами работы [10].

В корне иная ситуация наблюдается при использовании BIM (Basic Iterative Method) – итеративной атаки без проекции на L -шар. Несмотря на формальное сходство с PGD, отсутствие ограничения на максимальное изменение пикселя приводит к более «естественному» распределению возмущений.

Как видно из таблицы 3, для Model_A при $\epsilon = 0.2$ обнаруживаемость BIM падает более чем в 200 раз по сравнению с FGSM.

Особый интерес представляет сравнение эффективности обнаружения BIM на разных архитектурах. Как видно из таблицы 4, несмотря на более «сильную» архитектуру, Model_B показывает меньшее значение $TPR@FPR = 1\%$ против BIM.

В то же время Model_B проявляет наивысшую способность к обнаружению «грубых» атак (FGSM, PGD): при $\epsilon = 0.1$ метод PCA обеспечивает $TPR = 0.998$ при $FPR = 1\%$. Это объясняется тем, что глобальное усреднение в Model_B формирует признакововое пространство, в котором реконструкционная ошибка PCA четко отражает аномалии, вносимые крупномасштабными возмущениями.

Таблица 3. Падение эффективности обнаружения при переходе от FGSM к BIM (Model_A, $\epsilon = 0.2$)

Детектор	FGSM ($TPR@FPR = 1\%$)	BIM ($TPR@FPR = 1\%$)	Относительное снижение
OCVSM	0.972	0.004	-99.6 %

Таблица 4. Эффективность обнаружения BIM-атаки ($\epsilon = 0.2$) на базовых моделях

Модель	Лучший детектор	$TPR@FPR = 1\%$
Model_A	Mahalanobis	0.012
Model_B	PCA	0.004

4.2. Влияние силы состязательной атаки на эффективность обнаружения различных моделей. Если тип атаки определяет качественные особенности угрозы, то параметр силы возмущения (ϵ) задает ее количественную интенсивность. В рамках данного исследования, тестирование проводилось при трех значениях ϵ – 0.1, 0.15 и 0.2, – что соответствует максимальному смещению пикселя на 25.5, 38.25 и 51 единицу яркости (в исходном диапазоне [0, 255]). Такой диапазон охватывает как малозаметные ($\epsilon = 0.1$),

так и визуально различимые ($\varepsilon = 0.2$) возмущения, позволяя оценить, как изменяется эффективность детекции при переходе от скрытых (stealth) к явным атакам.

Анализ результатов выявляет три принципиально различных режима поведения, зависящих от сочетания архитектуры модели и наличия состязательного обучения.

Для незащищенных моделей (Model_A, Model_B) рост ε не приводит к значительному улучшению обнаруживаемости – особенно при строгих ограничениях на ложные срабатывания ($FPR \leq 1\%$). Рассмотрим динамику $TPR@FPR = 1\%$ для атаки BIM, представленную в таблице 5:

Таблица 5. Значения показателя $TPR@FPR$ при увеличении силы атаки

Модель	$\varepsilon = 0.1$	$\varepsilon = 0.15$	$\varepsilon = 0.2$	$\Delta(0.1 \rightarrow 0.2)$
Model_A (Mahalanobis)	0.010	0.010	0.012	+20 %
Model_B (PCA)	0.002	0.002	0.004	+100 %

Видно, что даже при удвоении ε с 0.1 до 0.2 рост TPR составляет всего 0.002-0.004, что делает детекторы практически бесполезными в реальных условиях, поскольку в большинстве систем информационной безопасности для эффективного обнаружения аномалий требуется значение $TPR \geq 0.8$ при $FPR \leq 0.01$.

Причина заключается в том, что базовые модели не обучены различать «тонкие» смещения в признаковом пространстве. Возмущения, даже при $\varepsilon = 0.2$, остаются в пределах «естественной» вариативности данных, и распределение активаций атакованных и чистых примеров сильно перекрывается.

Коренным образом иная картина наблюдается для моделей с состязательным обучением (Model_A_AT, Model_B_AT). Здесь рост ε приводит к резкому, нелинейному увеличению TPR, особенно при использовании метрики расстояния Махаланобиса.

Для Model_B_AT при атаке BIM значения показателя $TPR@FPR$ при увеличении силы атаки представлены в таблице 6:

Таблица 6. Зависимость показателя $TPR@FPR$ предобученной модели от силы атаки BIM

ε	AUC	$TPR@FPR = 1\%$	$TPR@FPR = 5\%$	$TPR@FPR = 10\%$
0.1	0.5806	0.054	0.194	0.240
0.15	0.6235	0.104	0.250	0.292
0.2	0.6758	0.164	0.286	0.312

Из данных таблицы следует, что при увеличении ϵ всего на 0.1 (от 0.1 до 0.2) $TPR@FPR = 1\%$ растет в 3 раза (с 0.054 до 0.164). Это говорит о том, что состязательное обучение формирует в признаковом пространстве четкие «зоны чувствительности», где даже небольшое превышение порога ϵ вызывает резкое изменение активаций – эффект, описанный в работе [23] как разделение состязательных подпространств (adversarial subspace separation).

Полученные результаты свидетельствуют о взаимном усилении между защитным обучением и применением расстояния Махаланобиса: защитное обучение формирует в признаковом пространстве компактные и четко разделимые классовые кластеры, а расстояние Махаланобиса, учитывающее ковариационную структуру этих кластеров, максимально чувствительно реагирует на отклонения, вносимые состязательными возмущениями.

Особый интерес представляет сравнение моделей Model_A_AT и Model_B_AT – систем, построенных с применением одинаковой схемы защитного обучения, но имеющих различную архитектуру сверточных слоев.

Для атаки BIM при $\epsilon = 0.2$ результаты, представленные в таблице 7, демонстрируют кардинальное различие в эффективности обнаружения:

Таблица 7. Эффективность обнаружения BIM-атаки ($\epsilon = 0.2$) моделями с защитным обучением, оцененная по расстоянию Махаланобиса

Модель	$TPR@FPR = 1\%$	$TPR@FPR = 5\%$	$TPR@FPR = 10\%$
Model A AT	0.010	0.058	0.124
Model B AT	0.164	0.286	0.312

Model_B_AT демонстрирует наиболее высокую обнаруживаемость базовых итеративных методов (BIM-атак) среди всех моделей. Так, при $\epsilon = 0.2$ расстояние Махаланобиса достигает $TPR = 0.164$ при $FPR = 1\%$, что в 16.4 раза выше, чем у Model_A_AT (0.010).

Данный эффект объясняется следующим образом:

- большое ядро (5×5) формирует более глобальные структурированные признаки, менее чувствительные к локальному шуму, но четко реагирующие на систематические смещения, вносимые атаками;
- формирование компактного 64-мерного пространства признаков (после операции глобального усреднения) минимизирует влияние «шумовых» измерений, что повышает соотношение сигнал/шум при вычислении расстояния Махаланобиса, поскольку

в низкоразмерных пространствах отклонения от центроида класса обладают большей статистической значимостью;

– оптимальный размер полносвязного слоя (512 нейронов) предотвращает избыточную выразительность модели, которая в противном случае могла бы интегрировать составительные примеры в свое распределение.

Следовательно, гиперпараметрическая оптимизация, изначально направленная на повышение точности классификации, косвенно способствовала лучшему обнаружению составительных атак.

Таким образом, сила атаки (ϵ) является критическим параметром при проектировании систем защиты, но ее влияние нелинейно и зависит от архитектуры и уровня защиты модели. Составительное обучение обеспечивает предсказуемое и управляемое поведение системы защиты: рост силы атаки приводит к стабильному увеличению вероятности ее обнаружения. Кроме того, предварительно проведенная гиперпараметрическая оптимизация нейронной сети, изначально направленная на повышение точности классификации, также способствует лучшему обнаружению составительных атак.

5. Заключение. В ходе исследовательской работы было проведено комплексное экспериментальное исследование, направленное на оценку эффективности методов обнаружения составительных атак на основе анализа аномалий в активациях промежуточных слоев сверточных нейронных сетей. Актуальность работы обусловлена растущими рисками, связанными с внедрением моделей машинного обучения в критически важные информационные системы, где устойчивость к целенаправленным атакам становится неотъемлемой частью требований к информационной безопасности.

Полученные результаты позволяют сделать следующие выводы:

1. Тип атаки определяет выбор метода защиты: FGSM и PGD легко обнаруживаются большинством детекторов ($TPR > 0.96$ при $FPR = 1\%$), в то время как BIM, благодаря своей «незаметной» структуре, снижает TPR до $0.004 - 0.012$ на базовых моделях – что подчеркивает необходимость тестирования на разнообразных атаках, а не только на FGSM.

2. Защитное обучение не только повышает устойчивость, но и усиливает способность к обнаружению атак, особенно в паре с методом обнаружения аномалий на основе расстояния Махаланобиса. Показано взаимное усиление между защитным обучением и статистической оценкой отклонений: первое формирует компактные, четко разделимые классовые кластеры, вторая – максимально чувствительна к их нарушению.

3. Архитектурные гиперпараметры оказывают решающее влияние на эффективность детекции. Модель Model_B, оптимизированная в предыдущем исследовании авторов [22] под критерий точности классификации (ядро 5×5 , глобальное усреднение, 512 нейронов в скрытом слое), продемонстрировала наивысшую способность к обнаружению «грубых» атак ($TPR = 0.998$ при $FPR = 1\%$ с использованием анализа главных компонент), что свидетельствует о совместимости задач повышения точности классификации и обеспечения безопасности информационных систем.

4. Метрика TPR при фиксированном $FPR = 1\%$ является более информативной, чем площадь под ROC-кривой (AUC), при проектировании систем защиты. Например, анализ главных компонент на Model_A показывает $AUC = 1.0$, но при фиксированном уровне ложноположительных срабатываний $FPR = 0.01$ достигается нулевая чувствительность ($TPR = 0$), что свидетельствует об отсутствии способности метода обнаруживать истинные аномалии в режиме, соответствующем практически приемлемой частоте ложных тревог.

Практическая значимость работы заключается в возможности интеграции полученных результатов в процессы обеспечения безопасности при разработке и эксплуатации информационных систем.

Литература

1. Rossolini G., Biondi A., Butazzo G. Attention-based real-time defenses for physical adversarial attacks in vision applications // Proceedings of the ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPs). 2024. pp. 23–32. DOI: 10.1109/ICCPs61052.2024.00009.
2. Kotenko I., Saenko I., Lauta O., et al. Analysis of Modern Research on Protection against Adversarial Attacks in Energy Systems // Informatics and Automation. 2025. vol. 24. no. 6. pp. 1751–1809. DOI: 10.15622/ia.24.6.8.
3. Wang H., Wu C., Zheng K. Defense against adversarial attacks based on color space transformation // Neural Networks. 2024. vol. 173. 106176 p. DOI: 10.1016/J.NEUNET.2024.106176.
4. Brau F., Rossolini G., Biondi A., Butazzo G. On the Minimal Adversarial Perturbation for Deep Neural Networks With Provable Estimation Error // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. vol. 45. pp. 5038–5052. DOI: 10.1109/TPAMI.2022.3195616.
5. Szegedy C., Zaremba W., Sutskever I., et al. Intriguing properties of neural networks // arXiv preprint arXiv:1312.6199. 2013.
6. Hong F., Bouhassira R., Chow J., et al. Comprehensive characterization of human color discrimination thresholds // bioRxiv. 2025. pp. 1–70. DOI: 10.1101/2025.07.16.665219.
7. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2015. pp. 1–11.
8. Papernot N., McDaniel P., Jha S., et al. The Limitations of Deep Learning in Adversarial Settings // Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P). 2016. pp. 372–387. DOI: 10.1109/EuroSP.2016.36.

9. Zhu Z., Chen H., Wang X., et al. Ge-AdvGAN: Improving the transferability of adversarial samples by gradient editing-based adversarial generative model // Proceedings of the 2024 SIAM International Conference on Data Mining (SDM). 2024. pp. 706–714.
10. Carlini N., Wagner D. Towards Evaluating the Robustness of Neural Networks // Proceedings of the IEEE Symposium on Security and Privacy (SP). 2017. pp. 39–57. DOI: 10.1109/SP.2017.49.
11. Wang Y., Liao X., Cui W., et al. Defending against and generating adversarial examples together with generative adversarial networks // Scientific Reports. 2025. vol. 15. 12994 p. DOI: 10.1038/s41598-024-83444-x.
12. Xu L., Zhai J. Generating adversarial examples with collaborative generative models // International Journal of Information Security. 2024. vol. 23. pp. 1077–1091. DOI: 10.1007/S10207-023-000780-1.
13. Madry A., Makelov A., Schmidt L., et al. Towards deep learning models resistant to adversarial attacks // Proceedings of the International Conference on Learning Representations (ICLR). 2018. pp. 1–28.
14. Kurakin A., Goodfellow I.J., Bengio S. Adversarial Examples in the Physical World // Proceedings of the 5th International Conference on Learning Representations. 2017. pp. 99–112.
15. Gui J., Sun Z., Wen Y., et al. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications // IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35. pp. 3313–3332. DOI: 10.1109/TKDE.2021.3130191.
16. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey // ACM Computing Surveys (CSUR). 2009. vol. 41. no. 3. pp. 1–58. DOI: 10.1145/1541880.1541882.
17. Li Z., Wang W., Li J., et al. UCG: A Universal Cross-Domain Generator for Transferable Adversarial Examples // IEEE Transactions on Information Forensics and Security. 2024. vol. 19. pp. 3023–3037. DOI: 10.1109/TIFS.2024.3352913.
18. Scholkopf B., Platt J.C., Shawe-Taylor J., et al. Estimating the Support of a High-Dimensional Distribution // Neural Computation. 2001. vol. 13. no. 7. pp. 1443–1471. DOI: 10.1162/089976601750264965.
19. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. 2006. vol. 27. no. 8. pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.
20. Lee K., Lee K., Lee H., Shin J. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks // Advances in Neural Information Processing Systems. 2018. vol. 31. pp. 1–11.
21. Feng W., Xu N., Zhang T., et al. Robust and Generalized Physical Adversarial Attacks via Meta-Gan // IEEE Transactions on Information Forensics and Security. 2023. vol. 19. pp. 1112–1125. DOI: 10.1109/TIFS.2023.3288426.
22. Клименко Д.М., Козлова Е.И. Исследование влияния гиперпараметров на точность нейросетевого предсказания с использованием набора данных Fashion-MNIST // Цифровая трансформация. 2026. Т. 32. №2. С. 44–52. DOI: 10.35596/1729-7648-2026-32-2-44-52.
23. Ma X., Li B., Wang Y., et al. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality // Proceedings of the 6th International Conference on Learning Representations (ICLR). 2018. pp. 1–15.

Козлова Елена Ивановна — канд. физ.-мат. наук, доцент, заведующий кафедрой, кафедра интеллектуальных систем, Белорусский государственный университет (БГУ). Область научных интересов: искусственный интеллект, машинное обучение, кибербезопасность. Число научных публикаций — 80. kozlova@bsu.by; ул. Курчатова, 5, 220047, Минск, Беларусь; р.т.: +375(17)209-5936.

Клименко Даниил Михайлович — магистрант, кафедра интеллектуальных систем, Белорусский государственный университет (БГУ). Область научных интересов: искусственный интеллект, машинное обучение, кибербезопасность. Число научных публикаций — 1. daniil.klimenko2212@yandex.by; ул. Курчатова, 5, 220047, Минск, Беларусь; р.т.: +375(17)209-5936.

Поддержка исследований. НИОКР № 20262086 от 08.07.2026 г. в рамках Задания 75 подпрограммы «Научно-учебное оборудование» государственной научно-технической программы «Научно-методическое и материально-техническое обеспечение образовательного процесса».

E. KOZLOVA, D.M. KLIMENKA
**DETECTING ADVERSARIAL ATTACKS THROUGH ANOMALY
IDENTIFICATION IN HIDDEN LAYERS OF A NEURAL
NETWORK**

Kozlova E., Klimenka D.M. Detecting Adversarial Attacks through Anomaly Identification in Hidden Layers of a Neural Network.

Abstract. The current level of information technology development is characterized by the widespread use of artificial intelligence (AI) methods, particularly deep machine learning, in such critical domains as healthcare, transportation, financial services, biometric authentication systems, and national security. Neural network models are increasingly employed to make decisions that directly affect the safety, health, and financial interests of individuals and organizations. However, as the functional significance of artificial intelligence systems increases, so does their vulnerability to targeted attacks. Existing countermeasures, such as adversarial training or input filtering, have significant limitations: they either degrade model accuracy on legitimate data or are effective only against previously known attack types, failing to detect emerging threats. Under these conditions, there is a growing need for detection mechanisms capable of identifying the very fact of an attack regardless of its implementation. Moreover, various studies have shown that adversarial examples induce anomalous activation patterns in the hidden layers of a neural network that differ qualitatively from those observed when processing legitimate data. This opens up the possibility of applying anomaly detection methods, such as isolation forest or Mahalanobis distance, to identify suspicious queries without modifying the original model. Of particular importance is also the influence of architectural and training hyperparameters on the internal representations of the model and, consequently, on its vulnerability and attack detection. The aim of this work is an experimental study of a method for detecting adversarial attacks based on anomaly analysis in the activations of hidden layers of a neural network.

Keywords: artificial intelligence, deep machine learning, neural networks, adversarial attack, attack detection, anomalies, hidden activations, internal representations of the model, isolation forest, model quality assessment metrics.

References

1. Rossolini G., Biondi A., Butazzo G., Attention-based real-time defenses for physical adversarial attacks in vision applications. Proceedings of the ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPs). 2024. pp. 23–32. DOI: 10.1109/ICCPs61052.2024.00009.
2. Kotenko I., Saenko I., Lauta O., et al. Analysis of Modern Research on Protection against Adversarial Attacks in Energy Systems. Informatics and Automation. 2025. vol. 24, no. 6. pp. 1751–1809. DOI: 10.15622/ia.24.6.8.
3. Wang H., Wu C., Zheng K. Defense against adversarial attacks based on color space transformation. Neural Networks. 2024. vol. 173. 106176 p. DOI: 10.1016/J.NEUNET.2024.106176.
4. Brau F., Rossolini G., Biondi A., Butazzo G. On the Minimal Adversarial Perturbation for Deep Neural Networks With Provable Estimation Error. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. vol. 45. pp. 5038–5052. DOI: 10.1109/TPAMI.2022.3195616.
5. Szegedy C., Zaremba W., Sutskever I., et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199. 2013.

6. Hong F., Bouhassira R., Chow J., et al. Comprehensive characterization of human color discrimination thresholds. *bioRxiv*. 2025. pp. 1–70. DOI: 10.1101/2025.07.16.665219.
7. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. 2015. pp. 1–11.
8. Papernot N., McDaniel P., Jha S., et al. The Limitations of Deep Learning in Adversarial Settings. *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*. 2016. pp. 372–387. DOI: 10.1109/EuroSP.2016.36.
9. Zhu Z., Chen H., Wang X., et al. Ge-AdvGAN: Improving the transferability of adversarial samples by gradient editing-based adversarial generative model. *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. 2024. pp. 706–714.
10. Carlini N., Wagner D. Towards Evaluating the Robustness of Neural Networks. *Proceedings of the IEEE Symposium on Security and Privacy (SP)*. 2017. pp. 39–57. DOI: 10.1109/SP.2017.49.
11. Wang Y., Liao X., Cui W., et al. Defending against and generating adversarial examples together with generative adversarial networks. *Scientific Reports*. 2025. vol. 15. 12994 p. DOI: 10.1038/s41598-024-83444-x.
12. Xu L., Zhai J. Generating adversarial examples with collaborative generative models. *International Journal of Information Security*. 2024. vol. 23. pp. 1077–1091. DOI: 10.1007/S10207-023-000780-1.
13. Madry A., Makelov A., Schmidt L., et al. Towards deep learning models resistant to adversarial attacks. *Proceedings of the International Conference on Learning Representations (ICLR)*. 2018. pp. 1–28.
14. Kurakin A., Goodfellow I.J., Bengio S. Adversarial Examples in the Physical World. *Proceedings of the 5th International Conference on Learning Representations*. 2017. pp. 99–112.
15. Gui J., Sun Z., Wen Y., et al. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering*. 2023. vol. 35. pp. 3313–3332. DOI: 10.1109/TKDE.2021.3130191.
16. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*. 2009. vol. 41. no. 3. pp. 1–58. DOI: 10.1145/1541880.1541882.
17. Li Z., Wang W., Li J., et al. UCG: A Universal Cross-Domain Generator for Transferable Adversarial Examples. *IEEE Transactions on Information Forensics and Security*. 2024. vol. 19. pp. 3023–3037. DOI: 10.1109/TIFS.2024.3352913.
18. Scholkopf B., Platt J.C., Shawe-Taylor J., et al. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*. 2001. vol. 13. no. 7. pp. 1443–1471. DOI: 10.1162/089976601750264965.
19. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006. vol. 27. no. 8. pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.
20. Lee K., Lee K., Lee H., Shin J. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *Advances in Neural Information Processing Systems*. 2018. vol. 31. pp. 1–11.
21. Feng W., Xu N., Zhang T., et al. Robust and Generalized Physical Adversarial Attacks via Meta-Gan. *IEEE Transactions on Information Forensics and Security*. 2023. vol. 19. pp. 1112–1125. DOI: 10.1109/TIFS.2023.3288426.
22. Klimenko D., Kozlova E. [Investigation of the Impact of Hyperparameters on the Accuracy of Neural Network Predictions Using the Fashion-MNIST Dataset]. *Cifrovaya transformaciya – Digital Transformation*. 2026. vol. 32. no. 2. pp. 44–52. DOI: 10.35596/1729-7648-2026-32-2-44-52. (In Russ.).

23. Ma X., Li B., Wang Y., et al. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality. Proceedings of the 6th International Conference on Learning Representations (ICLR). 2018. pp. 1–15.

Kozlova Elena — Ph.D., Associated Professor, Head of the Department, Department of Intelligent Systems, Belarusian State University (BSU). Research interests: artificial intelligence, machine learning, cybersecurity. The number of publications — 80. kozlova@bsu.by; 5, Kurchatov St., 220047, Minsk, Belarus; office phone: +375(17)209-5936.

Klimenka Daniil — Master's Student, Department of Intelligent Systems, Belarusian State University (BSU). Research interests: artificial intelligence, machine learning, cybersecurity. The number of publications — 1. daniil.klimenko2212@yandex.by; 5, Kurchatov St., 220047, Minsk, Belarus; office phone: +375(17)209-5936.

Acknowledgements. Research and Development Project No. 20262086 dated July 8, 2026, within the framework of Task 75 of the subprogram “Scientific and Educational Equipment” of the state scientific and technical program “Scientific, Methodological and Material-Technical Support of the Educational Process”.